Analyzing Hate Speech Data along Racial, Gender and Intersectional Axes

Anonymous ACL submission

Abstract

Warning: *This work contains strong and offensive language, sometimes uncensored.*

To tackle the rising phenomenon of hate speech, efforts have been made towards data curation and analysis. When it comes to analysis of bias, previous work has focused predominantly on race. In our work, we further investigate bias in hate speech datasets along racial, gender and intersectional axes. We identify strong bias against AAE, male and AAE+Male tweets, which are annotated as disproportionately more hateful and offensive than from other demographics. We provide evidence that BERT-based models propagate this bias and show that balancing the training data for these protected attributes can lead to fairer models with regards to gender, but not race.

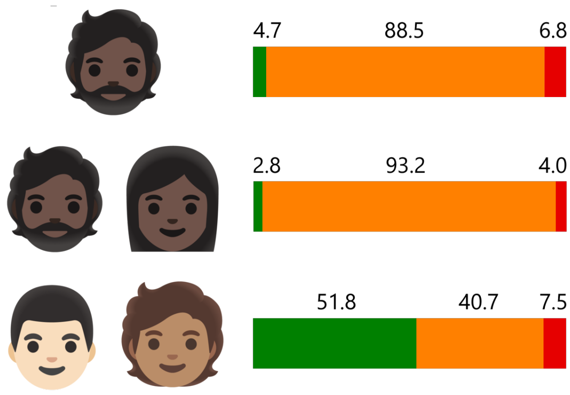

Figure 1: Distributions of label annotations on DAVIDSON (neutral, offensive, hateful) for AAE+Male, AAE and SAE (top-to-bottom). AAE has a higher ratio of offensive examples than SAE, while AAE+Male is both highly offensive and hateful.

1 Introduction

Hate Speech. To tackle the phenomenon of online hate speech, efforts have been made to curate datasets (Davidson et al., 2017; Guest et al., 2021; Sap et al., 2020). Since datasets in this domain are dealing with sensitive topics, it is of upmost importance that biases are kept to a (realistic) minimum and that data is thoroughly analyzed before use (Davidson et al., 2019a; Madukwe et al., 2020). In our work, we are contributing to this analysis by uncovering biases along the racial, gender and intersectional axes.

Racial, Gender and Intersectionality Biases. In data collection projects, biases can be introduced in a dataset due to–among other reasons–lack of annotator training or divergence between annotators and user demographics. For example, oftentimes the majority of annotators are white or male (Sap et al., 2020; Founta et al., 2018). An annotator not in the 'in-group' may hold (un)conscious biases based on misconceptions about 'in-group' speech which may affect their perception of speech from certain communities (O'Dea et al., 2015), leading to incorrect annotations when it comes to dialects the annotators are not familiar with. A salient example of this is annotators conflating African American English (AAE) with offensive or hateful language (Sap et al., 2019).

Intersectionality (Crenshaw, 1989) is a framework for examining how different forms of inequality (for example, racial or gender inequalities) intersect with and reinforce each other. These new social dynamics need to be analyzed both separately and as a whole in order to address challenges faced by the examined communities. For example, a black woman does not face inequality based only on race or only on gender: she faces inequality because of both these characteristics, separately and in conjunction. In this work, we are analyzing not only the racial or gender inequalities in hate speech datasets, but their intersectionality as well.

With research in the area of hate speech, the NLP community aims at protecting target groups and fostering a safer online environment. In this sensitive area, it is pivotal that datasets and models are analyzed extensively to ensure the biases we are protecting affected communities from do not appear in the data itself, causing further marginalization (for example, by removing AAE speech disproportionately more often).

Contributions. In summary, we **(i)** investigate racial, gender and intersectional bias in three hate speech datasets, Founta et al. (2018); Davidson et al. (2017); Mathew et al. (2021), **(ii)** examine classifier predictions on existing, general-purpose AAE/SAE and gendered tweets, **(iii)** identify model bias against AAE, male and AAE+Male (labeled as both AAE and male) tweets, **(iv)** show that balancing training data for gender leads to fairer models.

2 Related Work

Hate speech research has focused on dataset curation (Davidson et al., 2017; Founta et al., 2018; Sap et al., 2020; Guest et al., 2021; Hede et al., 2021; Grimminger and Klinger, 2021) and dataset analysis (Madukwe et al., 2020; Wiegand et al., 2019; Swamy et al., 2019). In our work, we further analyze datasets to uncover latent biases.

It has been shown that data reflects social bias inherent in annotator pools (Waseem, 2016; Al Kuwatly et al., 2020; Davidson et al., 2019a,b). Work has been conducted to identify bias against AAE (Sap et al., 2019; Zhou et al., 2021; Xia et al., 2020) and gender (Excell and Al Moubayed, 2021).

Kim et al. (2020) investigated whether bias along the intersectional axis exists in Founta et al. (2018). While Kim et al. (2020) focused on bias within a single dataset, in our work we generalize to multiple hate speech datasets. We also examine classifier behavior and methods to mitigate this bias.

Research from a sociolinguistic perspective has shown that genders exhibit differences in online text (Gefen and Ridings, 2005) as well as general speech (Penelope Eckbert, 2013). In Bamman et al. (2014) and Bergsma and Van Durme (2013), gender classifiers for English tweets were developed with accuracy of 88% and 85% respectively. In our work, we develop a gender classifier of tweets as well, focusing on precision over recall, leading to a smaller but more accurate sample of gendered data.

3 Datasets

Five English datasets were used: three hate speech datasets (DAVIDSON, FOUNTA and HATEXPLAIN), one dataset of tweets labeled for race (GROENWOLD) and one for gender (VOLKOVA).

	Neutral	Offensive	Hateful
DAVIDSON	0.95	0.95	0.42
FOUNTA	0.86	0.88	0.37
HATEXPLAIN	0.69	0.50	0.72

Table 1: F1-score of BERT for each label, evaluated on DAVIDSON, FOUNTA and HATEXPLAIN.

DAVIDSON. In Davidson et al. (2017), a hate speech dataset of tweets was collected, labeled for neutral, offensive and hateful language. Hateful language is defined as speech that contains expressions of hatred towards a group or individual on the basis of protected attributes like ethnicity, gender, race and sexual orientation.

FOUNTA. In Founta et al. (2018) a crowd-sourced dataset of tweets was presented, labeled for normal, abusive and hateful language. To unify definitions, we rename normal to neutral language and abusive to offensive language.

HATEXPLAIN. Mathew et al. (2021) presented a dataset from Twitter and Gab[1] passages. It has been labeled for normal (neutral), offensive and hateful language.

GROENWOLD. In Groenwold et al. (2020) a dataset of African American English and Standard American English tweets was introduced. The AAE tweets come from (Blodgett et al., 2016) and the SAE are direct translations of those tweets provided by annotators.

VOLKOVA. Volkova et al. (2013) presented a dataset of 800k English tweets from users with self-identified gender (female/male).

4 Experimental Setup

AAE Classifier. To classify tweets as AAE or SAE, we used the Blodgett et al. (2016) classifier. Since the number of tweets in our examined datasets was not sufficiently large, we could not use the recommended 0.8 threshold since it did not yield enough results for a confident analysis. We instead fell back to the 0.5 threshold, which can be interpreted as a straightforward classifier of AAE/SAE (whichever class has the highest score is returned).

Gender Classifier. To classify tweets as male or female, we finetuned BERT-base[2] on Volkova et al. (2013), which includes gender information as self-reported from the tweet authors. We split

[1] Gab is a social platform that has been known to host far-right groups and rhetoric.

[2] https://huggingface.co/bert-base-cased

	Male	Female	SAE	AAE	SAE+Male	SAE+Female	AAE+Male	AAE+Female
DAVIDSON	2716	2338	3534	8099	1279	1240	3157	1172
FOUNTA	26307	13615	43330	4177	13486	13257	971	787
HATEXPLAIN	4509	1103	10368	1103	4145	2376	250	240
GROENWOLD $_{AAE}$	586	613	0	1995	0	0	587	612
GROENWOLD $_{SAE}$	587	601	1980	0	587	601	0	0
VOLKOVA	41164	58836	37874	3755	16243	21631	1843	1912

Table 2: Protected attribute statistics for DAVIDSON, FOUNTA, HATEXPLAIN, GROENWOLD and VOLKOVA.

the dataset into train/dev/test (50K/25K/25K) and employed a confidence score of 0.8 as the threshold for assigning gender to a tweet. For the tweets with a confidence over the given threshold, precision was 78.4% when classifying tweets as 'male' and 79.5% when classifying tweets as 'female'.

Hate Speech Classifiers. For each of the three hate speech datasets (DAVIDSON, FOUNTA and HATEXPLAIN) we finetuned BERT-base. Performance is shown in Table 1. In DAVIDSON and FOUNTA, BERT performs well for neutral and offensive examples, but performs poorly for hateful content. In HATEXPLAIN, BERT overall performs worse, with slightly better performance for neutral and hateful examples over offensive ones.

Intersectionality. For our analysis, we classified tweets from all datasets for gender and race.

5 Intersectionality Statistics

In Table 2, we present statistics for gender, race and their intersection as found in the three examined hate speech datasets as well as in GROENWOLD and VOLKOVA. We show that no dataset is balanced between AAE and SAE. In FOUNTA and HATEXPLAIN, AAE tweets make up approximately 1/10th of the data. In DAVIDSON, we see stronger representation of AAE, with the AAE tweets being almost twice as many as the SAE tweets. DAVIDSON is also balanced for gender. The other hate speech datasets, while still not balanced, are more balanced for gender than they are for race. FOUNTA has twice as many male than female tweets and HATEXPLAIN has four times as many.

In Table 3, we present a breakdown of protected attributes per class (neutral/offensive/hateful) for DAVIDSON, FOUNTA and HATEXPLAIN. A main takeaway for DAVIDSON and FOUNTA is the imbalance of AAE versus SAE. In SAE, the neutral class makes up 52% of the data for DAVIDSON and 81% for FOUNTA, while the respective numbers for AAE are 3% for DAVIDSON and 13% for FOUNTA.

In HATEXPLAIN, AAE and SAE are more balanced, but there is instead imbalance between genders. For male and female speech, passages are neutral at rates of 43% and 61% respectively. In DAVIDSON, SAE+Female speech is viewed as more offensive than SAE+Male (48% vs. 19%), while in HATEXPLAIN, SAE+Male is more hateful than SAE+Female (34% vs. 16%). Finally, when comparing genders in AAE speech, we see that while AAE+Female contains a larger percentage of offensive tweets (for example, in FOUNTA, 69% vs. 54% and in HATEXPLAIN, 50% vs. 21%), AAE+Male contains disproportionately more hateful speech (in DAVIDSON, 7% vs. 5%, in FOUNTA, 28% vs. 9% and in HATEXPLAIN, 19% vs. 6%).

Overall, AAE and male speech is annotated as more offensive and hateful than SAE and female speech. Further analyzing AAE, AAE+Male is viewed as more hateful than AAE+Female.

6 Bias in BERT

We investigate to what extent data bias is learned by BERT. We compare our findings against a dataset balanced for race and gender, to examine whether balanced data leads to fairer models. Namely, we compare a randomly sampled with a balanced set the DAVIDSON dataset.[3] In the balanced set we sample the same number of AAE and SAE tweets (3000) and the same number of male and female tweets (1750). We also include 8000 neutral tweets without race or gender labels. For the randomly sampled set, for a fair comparison, we sampled the same number of tweets as the balanced set.[4] All sampling was stratified to preserve the original label distributions. Results are shown in Table 4.

In the randomly sampled set, there is an imbalance both for gender and race. For gender, while male tweets are more hateful (3% vs. 1%), female tweets are more offensive (71% vs. 63%). For race, AAE is marked almost entirely as offensive (94%), while SAE is split in neutral and offensive (53%

[3] FOUNTA and HATEXPLAIN were not considered for this study as they do not contain enough AAE examples to make confident inferences.

[4] Experiments were conducted with the entirety of the original dataset with similar results. They are omitted for brevity.

	Male			Female			SAE			AAE			SAE+Male			SAE+Female			AAE+Male			AAE+Female		
	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H
Davidson	32.2	61.9	5.9	27.7	69.5	2.8	51.8	40.7	7.5	2.8	93.2	4.0	77.0	19.4	3.6	50.0	47.8	2.3	4.7	88.5	6.8	6.8	88.0	5.2
Founta	81.2	12.3	6.4	71.0	25.0	4.0	80.5	14.6	4.9	13.2	69.2	17.6	86.9	7.6	5.5	86.2	11.4	2.4	18.3	53.8	27.9	21.8	69.4	8.8
HateXplain	43.0	23.7	33.3	60.7	24.6	14.8	38.3	26.7	35.0	45.6	39.1	15.3	41.6	24.0	34.4	58.9	25.1	16.0	59.4	21.3	19.4	44.4	50.0	5.6

Table 3: Distribution of protected attribute annotations for neutral/offensive/hateful (N/O/H) examples.

	Male			Female			SAE			AAE			SAE+Male			SAE+Female			AAE+Male			AAE+Female		
	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H	N	O	H
Random	33.8	63.2	3.0	27.7	71.2	1.1	53.1	40.5	6.4	4.9	94.1	1.0	77.3	19.3	3.4	45.6	53.2	1.2	6.4	91.2	2.4	3.0	94.3	2.7
Balanced	25.3	71.5	3.2	25.4	71.1	3.5	54.3	39.2	6.5	4.3	95.1	1.6	71.0	22.8	6.2	52.3	46.4	2.3	5.8	92.1	2.1	6.2	93.1	0.7

Table 4: Distribution of predictions for protected attributes on random and balanced datasets based on DAVIDSON. The balanced set is balanced on race (equal number of AAE and SAE tweets) and gender (equal number of female and male tweets). Shown are percentages for neutral/offensive/hateful (N/O/H) predictions.

	All	AAE
DAVIDSON	niggerize, subhuman, bastards, border, pigfucking, feminist, wetbacks, savages, wetback, jumpers	queer, negros, niggaz, racial, shittiest, wet, savage, skinned, darky, fags
FOUNTA	moron, insult, muslims, aggression, puritan, haters, arabs, coloured, ousted, pedophiles	white, killing, pathetic, nigga, slave, niggas, sells, hell, children, violent
HATEXPLAIN	towelhead, muzzrat, muscum, negresses, niggerette, niglets, musloid, niggerish, niggery, gorilla	spic, fuck, faggots, gorilla, towel, sandnigger, zhid, coons, rag, fowl

Table 5: Top 10 most contributing words for DAVIDSON, FOUNTA and HATEXPLAIN as computed with LIME for hateful predictions.

and 41%). In the SAE subset of tweets, there is an imbalance between genders, with SAE+Female being marked disproportionately more often as offensive than SAE+Male (54% vs. 19%).

6.1 Balanced Training

In Table 4, before balancing, 34% of male and 28% of female tweets are marked as neutral. After balancing, these rates are both at 25%. There is an improvement in the intersection of AAE and gender, with the distributions of AAE+Male and AAE+Female tweets converging. For SAE, SAE+Male and SAE+Female distributions converge too, although still far apart. Overall, balanced data improves fairness for gender but not for race, which potentially stems from bias in annotation.

6.2 Interpretability with LIME

In Table 5, we show the top contributing words for offensive and hateful predictions in DAVIDSON, FOUNTA and HATEXPLAIN. We see that for AAE, terms such as 'n****z' and 'n***a' contribute in classifying text as non-neutral even though the terms are part of African American vernacular (Rahman, 2012), showing that this dialect is more likely to be flagged. In non-AAE speech (which includes–but is not exclusive to–SAE), we see the n-word variant with the '-er' spelling appearing more often in various forms, which is correctly picked up by the model as an offensive and hateful term. On both sets, we also see other slurs, such as 'f*ggots', 'moron' and 'wetback' (a slur against foreigners residing in the United States, especially Mexicans) being picked up, showing the model does recognize certain slurs and offensive terms.

7 Conclusion

In our work, we analyze racial, gender and intersectional bias in hate speech datasets. We show that tweets from AAE and AAE+Male users are labeled disproportionately more often as offensive. We further show that BERT learns this bias, flagging AAE speech as significantly more offensive than SAE. We perform interpretability analysis using LIME, showing that the inability of BERT to differentiate between variations of the n-word across dialects is a contributing factor to biased predictions. Finally, we investigate whether training on a dataset balanced for race and gender mitigates bias. This method shows mixed results, with gender bias being mitigated more than racial bias. With our work we want to motivate further investigation in model bias not only for the usual gender and racial attributes, but also for their intersection.

8 Ethical Considerations

In our work we are dealing with data that can catalyze harm against marginalized groups. We do not advocate for the propagation or adoption of this hateful rhetoric. With our work we wish to motivate further analysis and documentation of sensitive data that is to be used for the training of models (for example, using templates from Mitchell et al. (2019); Bender and Friedman (2018)).

Further, while classifying protected attributes such as race or gender is important in analyzing and identifying bias, care should be taken for the race and gender classifiers to not be misused or abused, in order to protect the identity of users, especially those from marginalized demographics who are more vulnerable to hateful attacks and further marginalization. In our work we only predict these protected attributes for investigative purposes and do not motivate the direct application of such classifiers.

Finally, in our work we only focused on English and a specific set of attributes. Namely, we considered race (African American) and gender. This is a non-exhaustive list of biases and more work needs to be done for greater coverage of languages and attributes.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 184–190, Online. Association for Computational Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Shane Bergsma and Benjamin Van Durme. 2013. Using conceptual class attributes to characterize social media users. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Sofia, Bulgaria. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of EMNLP*.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019a. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019b. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

Elizabeth Excell and Noura Al Moubayed. 2021. Towards equal gender representation in the annotations of toxic language detection. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 55–65, Online. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*. AAAI Press.

David Gefen and Catherine Ridings. 2005. If you spoke as she does, sir, instead of the way you do: A sociolinguistics perspective of gender differences in virtual communities. *DATA BASE*, 36:78–92.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of EMNLP*.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

Anushree Hede, Oshin Agarwal, Linda Lu, Diana C. Mutz, and Ani Nenkova. 2021. From toxicity in online comments to incivility in American news: Proceed with caution. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2620–2630, Online. Association for Computational Linguistics.

Jae-Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *CoRR*, abs/2005.05921.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 220–229, New York, NY, USA. Association for Computing Machinery.

Conor J. O'Dea, Stuart S. Miller, Emma B. Andres, Madelyn H. Ray, Derrick F. Till, and Donald A. Saucier. 2015. Out of bounds: factors affecting the perceived offensiveness of racial slurs. *Language Sciences*, 52:155–164. Slurs.

Sally McConnell-Ginet Penelope Eckbert. 2013. *Language and Gender*. Cambridge University Press.

Jacquelyn Rahman. 2012. The n word: Its history and use in the african american community. *Journal of English Linguistics*, 40(2):137–171.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827, Seattle, Washington, USA. Association for Computational Linguistics.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Yejin Choi, and Noah Smith. 2021. Challenges in automated debiasing for toxic language detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3143–3155, Online. Association for Computational Linguistics.